Capstone Project - The Battle of Neighborhoods (Week 1)

1. Background

Assume an entrepreneur plans to open a fitness club in Los Angeles and she/he wants to know where she/he should choose to be the location of her/his club. There are totally 272 neighborhoods in LA. To simplify the analysis, we will reduce the number of neighborhoods to explore. Apparently, we want to have more people to become a member of the club and so firstly we want to consider the population factor of the neighborhoods. Therefore, we will only select the most ten populous neighborhoods to discuss.

2. Description of data

Data from three sources will be used:

1) All 272 neighborhoods with location coordinates, download from https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr

	Neighborhood	Longitude	Latitude
0	Acton	-118.169810	34.497355
1	Adams-Normandie	-118.300208	34.031461
2	Agoura Hills	-118.759885	34.146736
3	Agua Dulce	-118.317104	34.504927
4	Alhambra	-118.136512	34.085539

2) the top 10 high density neighborhoods in LA, retrieved from https://en.wikipedia.org/wiki/Template:High-density_neighborhoods_in_Los_Angeles_County

3) venues data of these ten neighborhoods explored from Foursquare API.

Firstly, the first two data sets will be combined to get coordinate information of the top10 neighborhoods for the later-on venue data extraction from Foursquare. Secondly, the ten neighborhoods will be clustered according to their similarity in venue categories. Lastly, the clustering results will be visualized after variable reduction using principle component analysis. After this analysis, we should be able to select some neighborhoods that are more suitable than others to have a new fitness club. If we are not be able to select, alternative clustering algorithms might need to apply.